Niveau(x) d'enseignement: 6 Numéro EDUQ : 8670192
Centre d'intérêt: 841

Bertrand, Richard G., 1949-.

L'analyse exploratoire des données comme approche pédagogique pour l'enseignement de la statistique / Richard Bertrand, Claude Valiquette.
-- Sainte-Foy : INRS-Education, 1982.
31 f. : graph. ; 28 cm.

Code : C-99.

Vedette(s)-secondaire(s): Valiquette, Claude * INRS-Education

Descripteur(s): Statistique * Mathématiques * Analyse de données * Pratique pédagogique * Collégial (Ordre d'enseignement) * Universitaire (Ordre d'enseignement)

Identificateur(s) de sujet: Analyse exploratoire des données * Analyse confirmatoire des données

Dans cet article, les auteurs spécifient d'abord le rôle de l'informatique dans l'analyse des données, en distinguant analyse des données et traitement des données; l'ordinateur, utile pour le traitement des données, n'est pas indispensable pour l'analyse même des données. Ils abordent ensuite l'objet même de leur étude: la distinction qu'il faut faire entre approche exploratoire et approche confirmatoire dans l'analyse des données. Ils traitent de l'importance des graphes en analyse exploratoire, puis du diagramme en feuille et du diagramme en boîte.

Type de document: r

EDUQ 8670192

L'ANALYSE EXPLORATOIRE DES DONNEES COMME APPROCHE PEDAGOGIQUE POUR L'ENSEIGNEMENT DE LA STATISTIQUE.

Richard Bertrand
Claude Valiquette

C-99

INRS-Education Novembre 1982

EDUQ 8670192

L'ANALYSE EXPLORATOIRE DES DONNÉES COMME APPROCHE PÉDAGOGIQUE POUR L'ENSEIGNEMENT DE LA STATISTIQUE

Richard Bertrand, prof., INRS-Éducation
Claude Valiquette, prof., Université Laval

C-99

INRS-Éducation
2383, Chemin Sainte-Foy
Ste-Foy, Québec
G1V 1T1

Novembre 1982

L'analyse exploratoire des données comme approche pédagogique pour l'enseignement de la statistique

I) INTRODUCTION

L'analyse des données est une expression de plus en plus répandue, et ce surtout dans les milieux scientifiques, milieux où l'on fait de la recherche. Dans les milieux scolaires secondaire, collégial ou universitaire, on connaît moins cette expression. L'utilisation du terme générique "statistique" que l'on associe d'ailleurs abondamment à "probabilité", est plus courante.

A l'instar du célèbre John W. Tukey, nous préférons utiliser la première expression plus explicite et plus pertinente que la seconde surtout si on veut appliquer plutôt que théoriser (comme c'est le lot de la plupart).

Indiquons d'abord ce que l'on entend ici par "donnée": c'est une pièce d'information quantitative recueillie en vue de l'étude d'un phénomène. Par exemple, l'ensemble suivant {120, 125, 138, 119, 145, 135, 160, 151, 143, 130} constitue un paquet de données concernant la grandeur en cm de 10 étudiants de Sec. I. Celui-ci {10, 48, 32, 25, 51, 20, 5, 0} indique la fréquence du nombre de questions posées par 8 enseignants pendant une période de 5 minutes. Ce dernier {-30, -21, -28, -13, -24, -7, -2, 0, 4, -8, -12, -26} nous montre la température en degrés Celsius (prise à 13h00) pour un échantillon de 12 jours au mois de janvier 199?

Au cours de cette communication nous emploierons donc l'expression "analyse des données" comme signifiant:

1) l'expression visuelle ou géométrique des données (graphique, tabulaire ou autre),
2) la recherche d'indices qui résument le plus précisément possible comment les données sont distribuées: mesures de tendance centrale, de dispersion, d'asymétrie,
3) l'étude des regroupements de données,
4) la détection des valeurs aberrantes et la ré-expression (transf. log., exp. ou autre) des données,
5) l'établissement d'hyp. stat. pertinentes,
6) le choix des méthodes d'analyse
7) les tests d'hypothèses,
8) l'estimation de paramètres (ponctuelle ou à l'aide d'intervalles de confiance).

Avant de pousser plus à fond l'étude de "l'analyse des données", nous aimerions bien spécifier le rôle de l'informatique en ce domaine.

II) L'INFORMATIQUE ET L'ANALYSE DES DONNEES

Depuis quelques années, il paraît virtuellement irréfutable d'avouer que l'informatique a joué un rôle fort important, tout au moins quantitativement, en analyse des données.

En particulier pour les tests d'hypothèses et l'estimation, son rôle a été prépondérant. Cependant un trop grand engouement pour les logiciels "servant à toutes les sauces" a joué un mauvais tour aux utilisateurs. On a tôt fait de surestimer l'importance de l'ordinateur en analyse de données. C'est devenu une panacée. Tous nos problèmes d'analyse se réglaient du coup. Il est même pratique courante d'utiliser une méthode d'analyse plutôt qu'une

autre (peut-être plus pertinente) sur la seule base de sa disponibilité dans un logiciel quelconque. Il s'agit là d'un cas typique d'une mauvaise utilisation de l'ordinateur. De même que c'est le sabbat qui est fait pour l'homme et non l'inverse, l'ordinateur est fait par et pour l'homme (la machine au service de l'homme) et non pas l'inverse (l'homme au service de la machine).

Il faut donc distinguer l'analyse et le traitement des données. De même que l'ordinateur est au service de l'homme, le traitement des données est au service de l'analyse, au même titre que le papier et le crayon. D'ailleurs l'analyse des données existait bien avant l'apparition de l'ordinateur souligne Tukey (1977, p. 663). Mais qu'entend-t-on par "traitement informatique des données"? Ca consiste essentiellement à utiliser des programmes informatiques:

1) soit pour calculer des formules mathématiques complexes et itératives, choisies au préalable par l'analyste dans le but de lui permettre d'accélérer son travail et ce en minimisant les erreurs de calcul,
2) soit pour représenter graphiquement des paquets de données dans le but d'en obtenir une image bien précise.

Il s'agit donc d'un travail (clérical) qui peut-être fort précieux et parfois même crucial pour l'analyste. Il serait cependant fort peu valorisant pour une personne de considérer qu'il s'agit là de l'essence même de son travail. Non le traitement est une fonction connexe de l'analyse tout comme l'ordinateur (et donc les logiciels) est un outil dans la main de l'analyste.

C'est pourquoi, contrairement à une pratique (malheureusement) populaire, il faut être très prudent lors de l'entrée des données pour fins de traitement. Un programme de validation étanche devrait

être de mise pour éliminer au moins les erreurs les plus flagrantes. Souvenons-nous de l'expression "Garbage in, garbage out".

Cet exposé concernant les distinctions à apporter entre l'analyse et le traitement des données a pu sembler un réquisitoire contre l'utilisation de l'ordinateur. Il n'en est rien. On a simplement voulu sensibiliser les auditeurs aux mauvaises utilisations de cet outil et plus spécialement des logiciels informatiques. Il s'agissait de remettre les choses à leur place afin d'éviter de "débrancher notre cerveau en branchant l'ordinateur"[1].

III APPROCHES EXPLORATOIRE ET CONFIRMATOIRE

La popularité de l'analyse des données telle que nous l'avons définie est due en bonne partie à un mathématicien du nom de John W. Tukey. Pour bien comprendre sa philosophie, citons un texte[2] de ce valeureux analyste des données qui remonte à plus de vingt ans:

> "Pendant vingt ans j'ai cru que j'étais un statisticien qui s'intéressait aux inférences allant du particulier au général. Mais attentif au développement de la statistique mathématique, j'ai trouvé des raisons d'étonnement et de doute".

Il n'en faut pas plus pour justifier deux approches complémentaires d'analyse des données: l'approche exploratoire et l'approche confirmatoire. Il peut être indiqué que cette typologie

---

1 "When me plug in the computer, we often unplug our brains" de Erickson, B.H. et T.A. Nosanchuk (1977), "Understanding data", McGraw-Hill, p. 28.

2 Ce texte est repris par G. Morlat dans Cailliez, F. et J.P. Pages, "Introduction à l'analyse des données" SMASH, 1976.

"exploratoire-confirmatoire" se rapproche un peu de celle plus connue "descriptive-inférentielle". Mais avec la différence toutefois que les techniques exploratoires mises de l'avant par Tukey sont plus puissantes, plus visuelles et plus engageantes que les classiques statistiques descriptives.

C'est d'ailleurs l'objet de cette communication de les présenter à l'aide d'exemples relevant d'un domaine d'application riche et motivant. Mais avant d'entrer tout de go dans ce sujet, nous aimerions emprunter une analogie due à Tukey pour bien préciser les fonctions distinctives de l'approche exploratoire et de l'approche confirmatoire.

Quittons donc un peu le monde des données statistiques pour entrer dans le domaine judiciaire. Tukey ne nous fait pas languir longtemps (à la manière par exemple d'Agatha Christie) pour nous faire sentir les similitudes entre l'A.E.D. et l'enquête policière où le détective passe le plus clair de son temps à rechercher des indices (pour prendre à faux le suspect, à la manière par exemple du Lieutenant Colombo). Il existe de la même façon des ressemblances frappantes entre l'A.C.D. et un procès à la cour de justice[1] où les avocats tentent de prouver l'innocence ou la culpabilité d'un suspect (à la manière par exemple de Perry Mason) en faisant les présomptions entre autres que le juge est impartial, les membres du jury sont normaux, sans biais et homogènes (ont la même importance aux fins du procès).

Tout comme au procès l'appareil judiciaire (le juge, les avocats et le jury) ne peut rien sans indices, sans pièces à con-

---

1 Voir Erickson, B.H. et T.A. Nosanchuk (1977, p. 145), op. cit.

viction, la phase d'A.C.D. est infructueuse sans indications. Dans un cas (analyse des données) comme dans l'autre (domaine judiciaire) il faut dans une première phase trouver des indices, un ou des suspects (une ou des hypothèses), avant de faire le procès (de vérifier une hypothèse) même si ce suspect est supposé innocent dès le point de départ (même si l'hypothèse est supposée "nulle"), compte tenu des présomptions de normalité des membres du jury (ou des populations statistiques) et d'homogénéité des membres du jury (ou des variances des populations statistiques).

Ce qui n'exclut pas l'interaction constante entre les deux phases de l'analyse (ou de l'enquête). De nouveaux indices pourraient permettre de vérifier d'autres hypothèses (ou d'inculper d'autres suspects), ce qui nous amènerait à d'autres indices puis de nouvelles hypothèses, etc...

Voici d'ailleurs, au tableau 1, une synthèse des comparaisons entre l'A.E.D. et l'A.C.D.

A ces distinctions, ajoutons que l'A.E. se prête, selon nous, beaucoup plus facilement que l'A.C. à un premier enseignement de l'analyse statistique des données et ce, surtout pour des étudiants non-matheux. La logique de l'A.E. est beaucoup plus intuitive, les concepts sont relativement simples et les méthodes n'exigent pas de présomptions fortes comme c'est le cas en A.C.

IV IMPORTANCE DES GRAPHES EN A.E.

Avant d'entrer à fond dans les techniques exploratoires comme telles, il convient d'étudier sommairement le rôle des graphes en A.E.D.

Exploratoire	Confirmatoire
. Approche descriptive	. Approche inférentielle
. Logique simple	. Logique rigide
. Indication des hypothèses à tester	. Test d'hypothèse
. Utilisation d'indices stat. résistants	. Utilisation d'indices moins résistants
. Plan de recherche souple	. Plan de recherche rigoureux
. Usage des données en mains	. Usage de données exemptes d'erreur (si possible)
. Importance de l'expression visuelle ou graphique des données	. Moindre importance de l'expression visuelle ou graphique
. Avant une phase confirmatoire	. Après une phase exploratoire
. Vision intuitive des données	. Vision précise des données
. Esprit ouvert, divergent	. Esprit fermé, convergent
. Analogie avec une enquête policière	. Analogie avec un procès
. Importance du jugement du chercheur	. Importance du jugement du consultant
. Moindre importance de l'ordinateur	. Grande importance de l'ordinateur

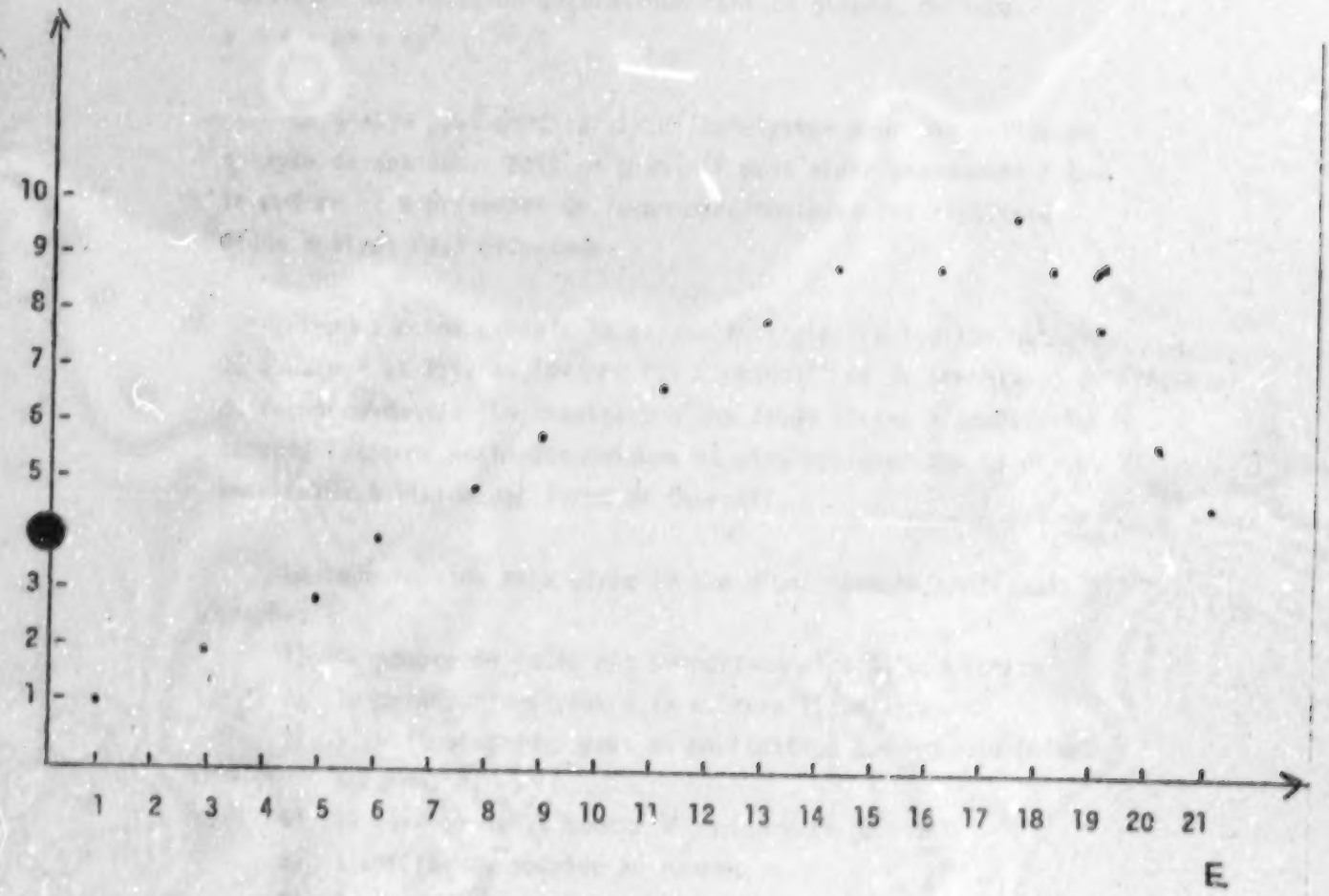
Tableau 1 Principales distinctions entre A.E.D. et A.C.D.

Regardons plutôt les scores suivants.

Étudiant / Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
E	1	3	5	6	8	9	11	13	14	16	17	18	19	20	21
O	1	2	3	4	5	6	7	8	9	9	10	9	8	6	5

Il s'agit du résultat de 15 étudiants à un test écrit (E) et à un test oral (O).

Un simple calcul du coefficient de corrélation de Pearson donne r = 0,75. Ce qui laisse croire à une relation linéaire relativement forte entre les deux séries de scores. Cependant la réalité est un peu moins simple. Regardons le graphe cartésien (aussi appelé correlogramme, dans ce cas) de cette relation (graphe 1).



GRAPHE 1 Exemple de graphe cartésien (relations linéaire et quadratique)

Un coup d'oeil, même rapide, à ce graphe, suffit pour nous convaincre qu'autre chose qu'une relation linéaire est en jeu. En fait, l'oeil connaisseur de l'analyste n'aurait aucun mal à diagnostiquer une relation quadratique dans ce graphe, du type $y = a + bx + cy^2$.

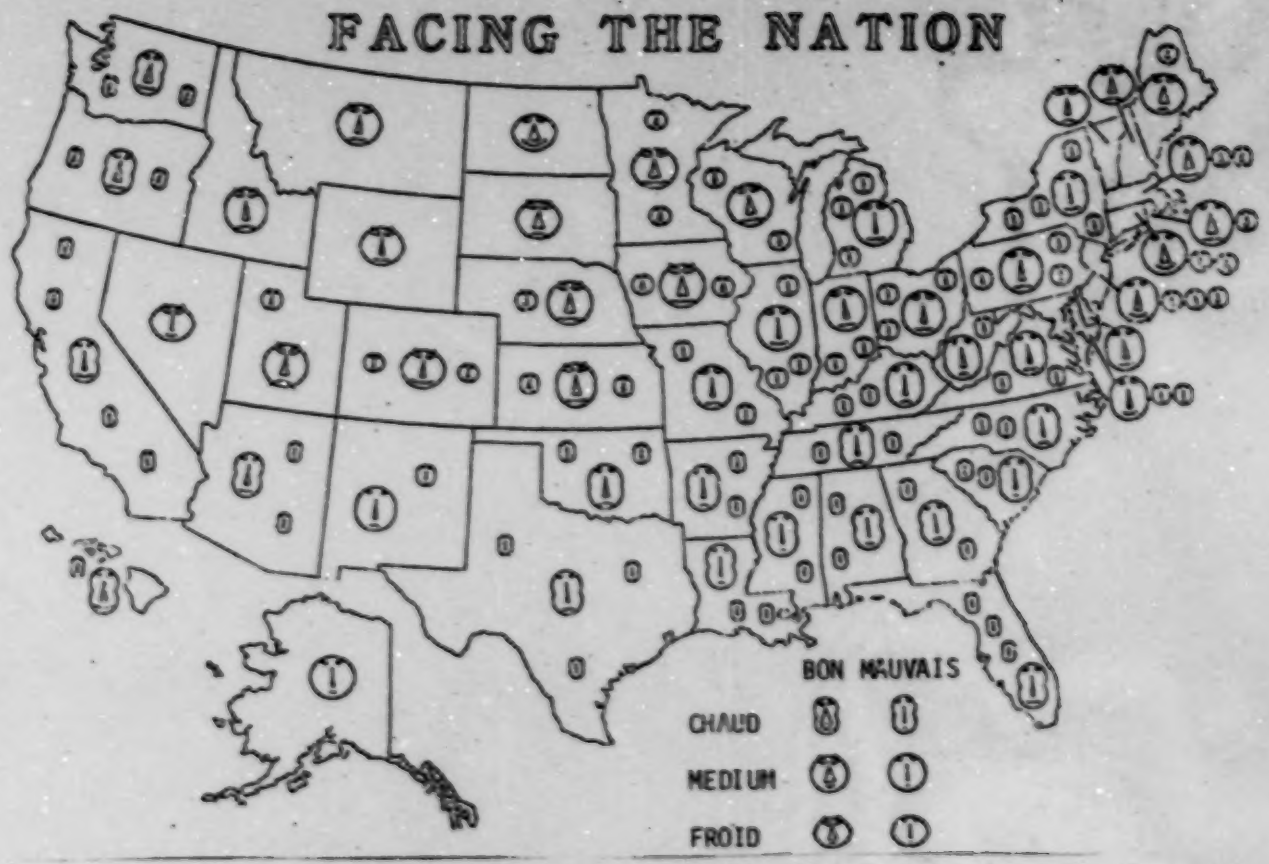
Un graphe peut donc servir de catalyseur pour une meilleure analyse de données. Mais en plus, il peut aider grandement à interpréter et à présenter de façon parcimonieuse les résultats d'une analyse déjà effectuée.

Prenons comme exemple le graphe intitulé "Facing the nation" dû à Wainer et Thissen (graphe 2). L'objectif ce de graphe est de présenter de façon condensée les résultats d'une étude visant à établir les caractéristiques socio-économiques et géographiques des 50 états américains à l'aide des faces de Chernoff.

Le code suivant sera utile en vue d'une lecture profitable du graphe:

1) le nombre de faces est proportionnel à la population
2) la grandeur des yeux à la culture littéraire
3) l'inclinaison des yeux au pourcentage des gradués (niveau secondaire)
4) la largeur de la bouche à l'espérance de vie
5) l'ampleur du sourire au revenu
6) la largeur du nez au taux de non-homicide
7) la largeur de la tête au nombre de jours froids.

Est-il possible de rassembler tous ces éléments en même temps sur un graphe! Il s'agit là d'un tour de force sans précédent. Remarquez aussi la clarté et le détail des caractéristiques, des physionomies.



GRAPHE 2 Diagramme des faces de Chernoff concernant les principales caractéristiques socio-économiques et géographiques des 50 états américains

En terminant cette section, nous donnons une série de règles à suivre en vue d'améliorer la présentation et donc de faciliter l'exploration d'un tableau de données:

- . arrondir suffisamment
- . ordonner les lignes et colonnes en utilisant les caractéristiques des données
- . espacer les données pour une meilleure perception
- . ajouter les indices statistiques sommaires pertinents.

Les tableaux 2 et 3 présentent les mêmes données, la différence résidant dans le fait que pour le tableau 2 on n'a pas suivi les règles prescrites. Avec lequel des tableaux aimeriez-vous le mieux travailler?

A part le fait que l'on passe de 1390 symboles (chiffres, points, signes de négation) pour le tableau 2, à 235 pour le tableau 3, il est à remarquer que la redondance a été éliminée dans ce dernier tableau.

De plus on a arrondi, enlevé les ".", ordonner les coefficients du plus haut au plus bas dans la première colonne tout en diminuant le nombre de pages de 2 à 1 et ce sans véritable perte d'information.

V LE DIAGRAMME EN FEUILLES ET LE DIAGRAMME EN BOITE

Souvenons-nous des représentations visuelles de données à l'aide des diagrammes classiques.

Tout d'abord le diagramme en bâtons, bien connu, offre cet aspect (graphe 3):

TABLEAU 2

Matrice de corrélations issue de SPSS

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
V_1	1.00000	.57548	.35737	.74084	.55100	.44292	.78165	.61892	.62517	.40994
V_2	.57548	1.00000	.61270	.95828	.67457	.76002	.90288	.67696	.94366	.59470
V_3	.35737	.61270	1.00000	.59518	.96110	.94610	.56783	.90704	.59113	.93268
V_4	.74084	.95828	.59518	1.00000	.71703	.75419	.98677	.75317	.97343	.62259
V_5	.55100	.67457	.96110	.71703	1.00000	.95961	.71542	.98679	.69843	.95835
V_6	.44292	.76002	.94610	.75419	.95961	1.00000	.72980	.93546	.77728	.95610
V_7	.78165	.90288	.56783	.98677	.71542	.72980	1.00000	.77067	.96170	.62150
V_8	.61892	.67696	.90704	.75317	.98679	.93546	.77067	1.00000	.73352	.94517
V_9	.62517	.94366	.59113	.97343	.69843	.77728	.96170	.73352	1.00000	.63956
V_{10}	.40994	.59470	.93268	.62259	.95835	.95610	.62150	.94517	.63956	1.00000
V_{11}	.44530	.76654	.87295	.77513	.91249	.98084	.76058	.90747	.82485	.93345
V_{12}	.36227	.61710	.90491	.62682	.92112	.96195	.61723	.90435	.66702	.98599
V_{13}	.07685	-.61321	-.37167	-.45759	-.30113	-.39967	-.37462	-.25749	-.44176	-.27387
V_{14}	.20250	-.27614	-.69701	-.17235	-.55849	-.55754	-.12677	-.47432	-.21021	-.50371
V_{15}	-.09750	.72714	.43895	.53178	.34936	.52736	.43024	.29649	.58613	.35555

TABLEAU 2 (suite)

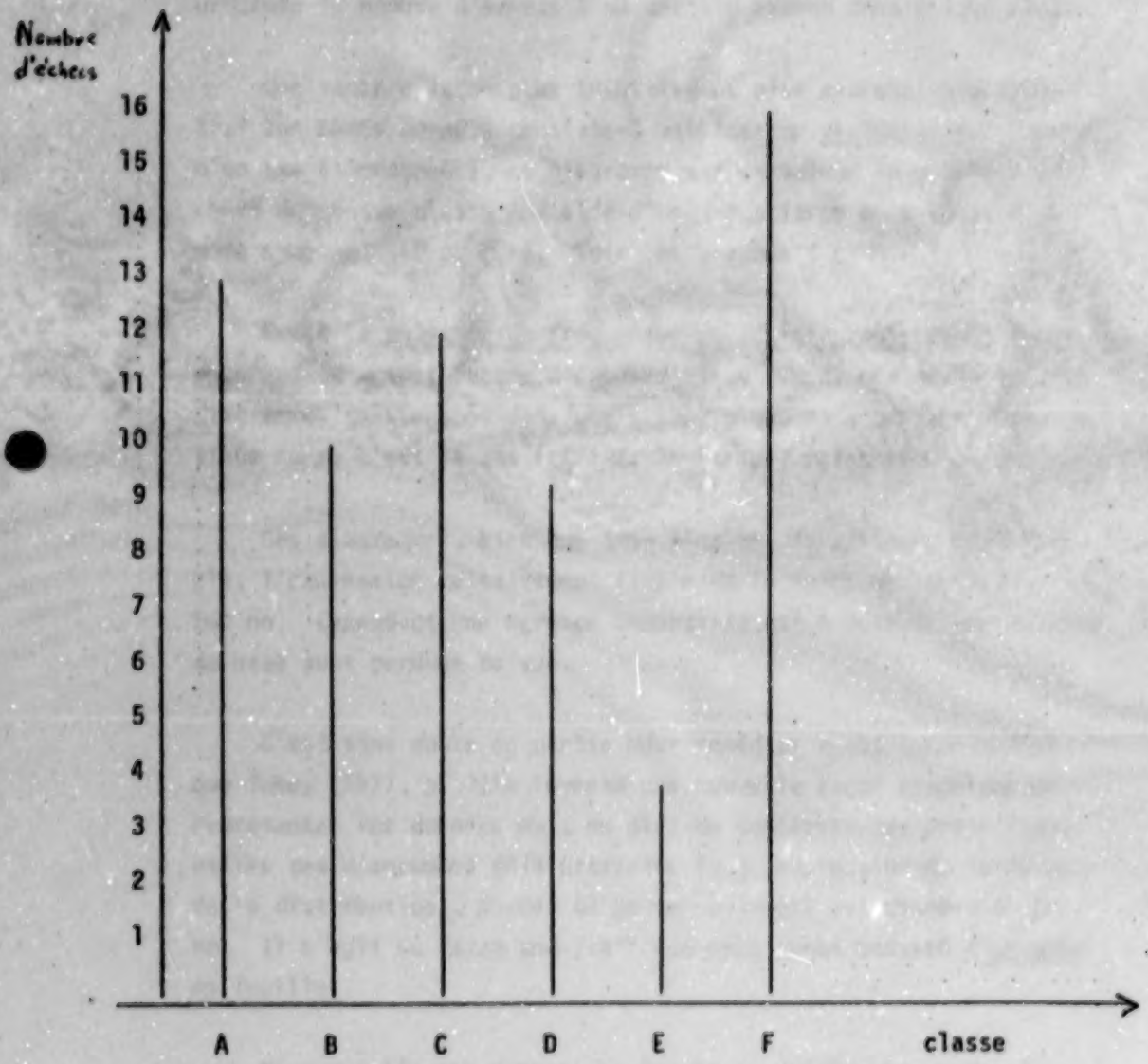
Matrice de corrélations issue de SPSS

	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}
V_1	.44530	.36227	.07685	.20250	-.09750
V_2	.76654	.61710	-.61321	-.27614	.72714
V_3	.87295	.90491	-.37167	-.69701	.43895
V_4	.77513	.62682	-.45759	-.17235	.53178
V_5	.91249	.92112	-.30113	-.55849	.34936
V_6	.98084	.96195	-.39967	-.55754	.52736
V_7	.76058	.61723	-.37462	-.12677	.43024
V_8	.90747	.90435	-.25749	-.47432	.29649
V_9	.82485	.66702	-.44176	-.21021	.58613
V_{10}	.93345	.98599	-.27387	-.50371	.35555
V_{11}	1.00000	.96075	-.36095	-.46579	.51921
V_{12}	.96075	1.00000	-.29072	-.47720	.41454
V_{13}	-.36095	-.29072	1.00000	.49612	-.80913
V_{14}	-.46579	-.47720	.49612	1.00000	-.48200
V_{15}	.51921	.41454	-.80913	-.48200	1.00000

TABLEAU 3

Matrice de corrélations transformée

	V_1	V_7	V_4	V_9	V_8	V_2	V_5	V_{11}	V_6	V_{10}	V_{12}	V_3	V_{14}	V_{13}
V_7	78													
V_4	74	99												
V_9	63	96	97											
V_8	62	77	75	73										
V_2	58	90	96	94	68									
V_5	55	72	72	70	99	67								
V_{11}	45	76	78	82	91	77	91							
V_6	44	73	75	78	94	76	96	98						
V_{10}	41	62	62	64	95	59	96	93	96					
V_{12}	36	62	63	67	90	62	92	96	96	99				
V_3	36	57	60	59	91	61	96	87	95	93	90			
V_{14}	20	-13	-17	-21	-47	-28	-56	-47	-56	-50	-47	-70		
V_{13}	07	-37	-46	-44	-26	-61	-30	-36	-40	-27	-29	-37	50	
V_{15}	-10	43	53	59	30	73	35	52	53	36	41	44	-48	-81



GRAPHE 3 Diagramme en bâtons du nombre d'échecs à un examen de statistiques dans six classes de Secondaire V

En abscisse on retrouve différents niveaux d'une variable (discrète) et en ordonnée une fréquence quelconque associée à cette variable.

Par exemple ici, en abscisse on retrouve des classes et en ordonnée le nombre d'échecs à un certain examen dans chaque classe.

Une seconde façon plus intuitive et plus expressive d'illustrer les mêmes données consiste à utiliser un pictogramme. Dégagé d'un axe (l'ordonnée), ce diagramme met en valeur le nombre d'échecs de chaque classe à l'aide d'une répétition en colonne d'un même symbole. (E ou ☹). Voir les graphes 4 et 5.

Enfin le polygone de fréquences et l'histogramme constituent encore deux autres façons d'exposer visuellement des données. Ces diagrammes conviennent dans une situation où on a une variable continue comme c'est le cas ici avec le "temps" (graphes 6 et 7).

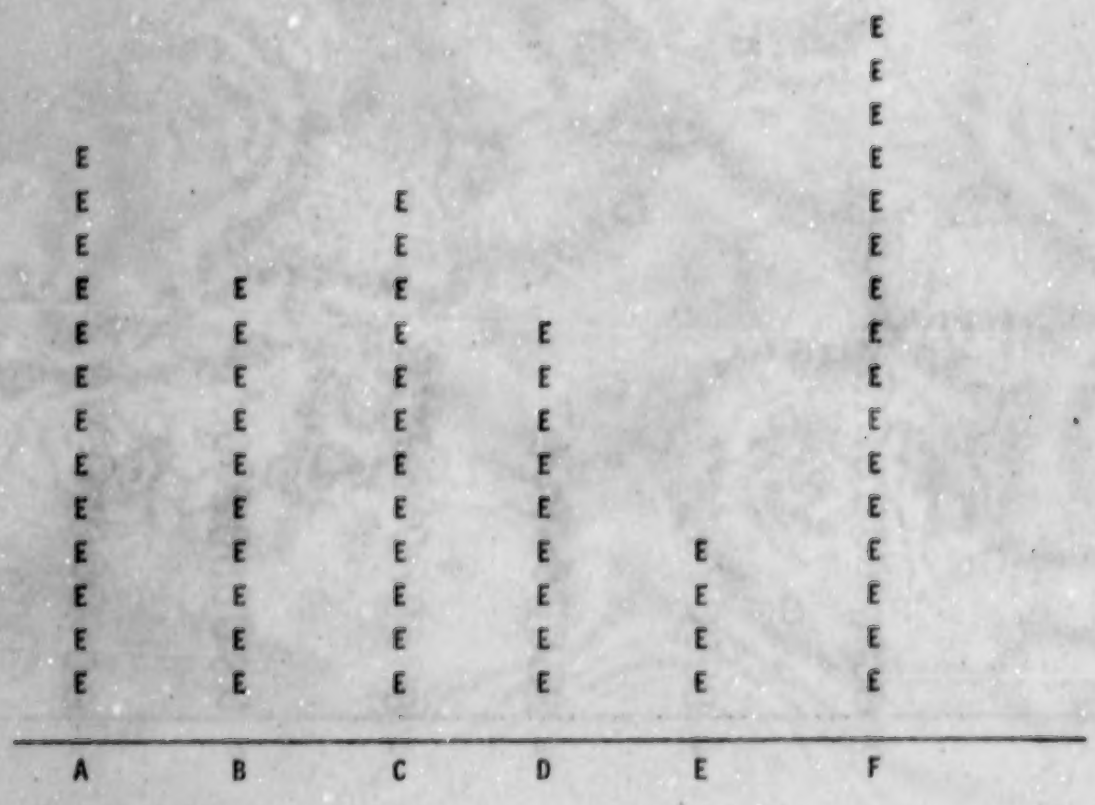
Ces diagrammes, bien que très simples, favorisent, par exemple, l'expression relativement fidèle de la forme de la distribution. Cependant une carence importante est à noter: les données de base sont perdues de vue.

C'est sans doute en partie pour remédier à cet inconvénient que Tukey (1977, p. 7)[1] a inventé une nouvelle façon graphique de représenter les données qui, en plus de conserver les propriétés utiles des diagrammes déjà présentés (v.g. expression de la forme de la distribution), permet de garder à l'oeil les données de base. Il s'agit du "stem-and-leaf" que nous avons baptisé diagramme en feuilles.

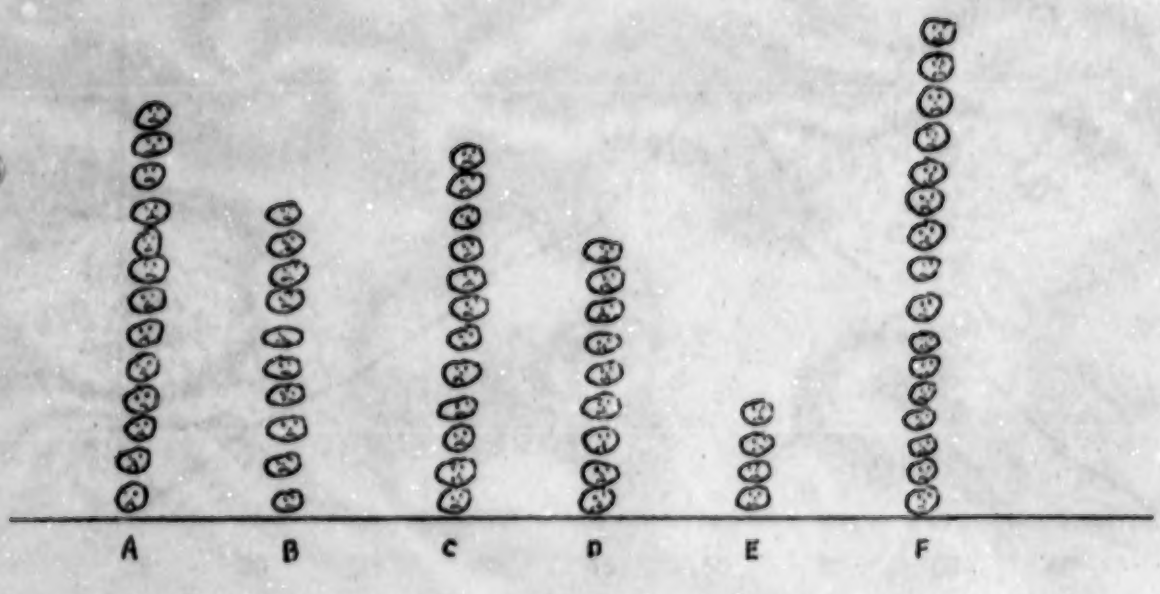
Nous appellerons branche ou "stem" le chiffre (ou nombre) situé à gauche de la barre verticale de chaque ligne.

---

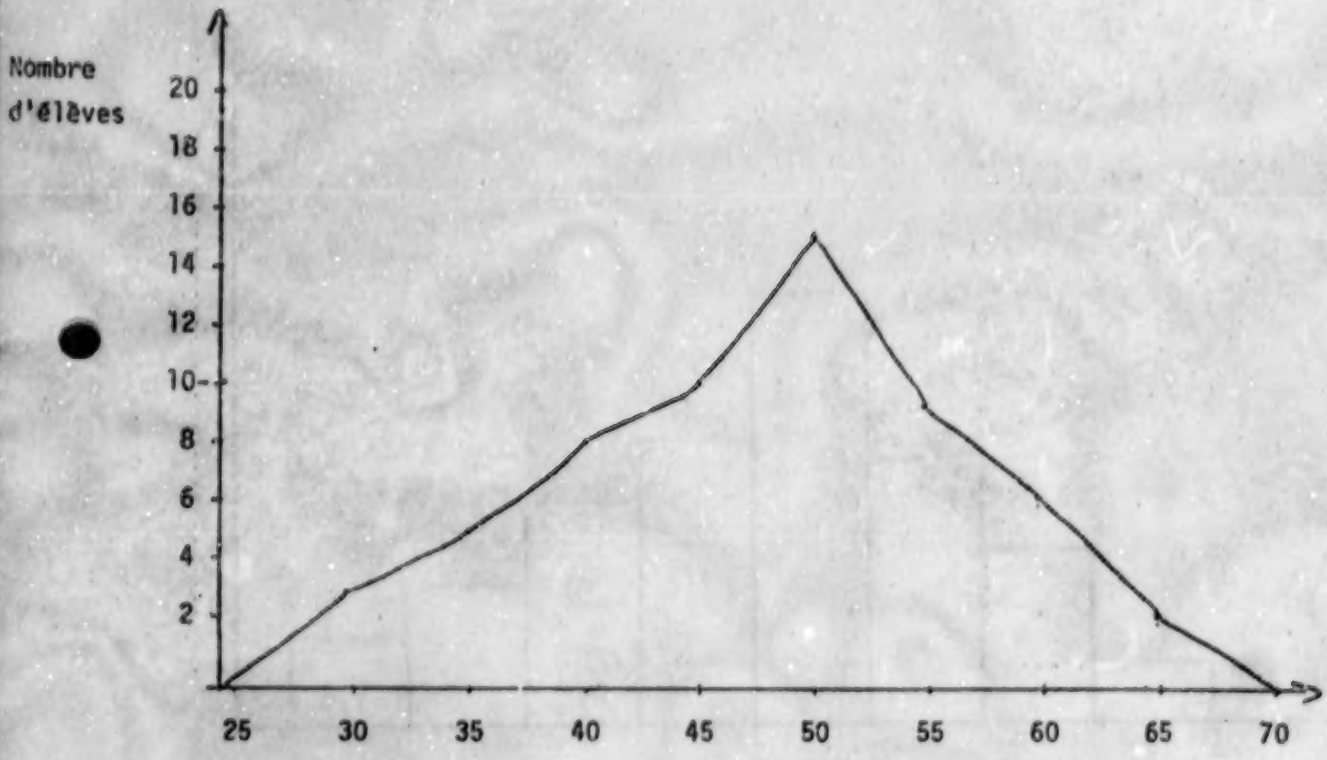
1 Tukey, J.W. "Exploratory data analysis", Addison-Wesley, 1977.



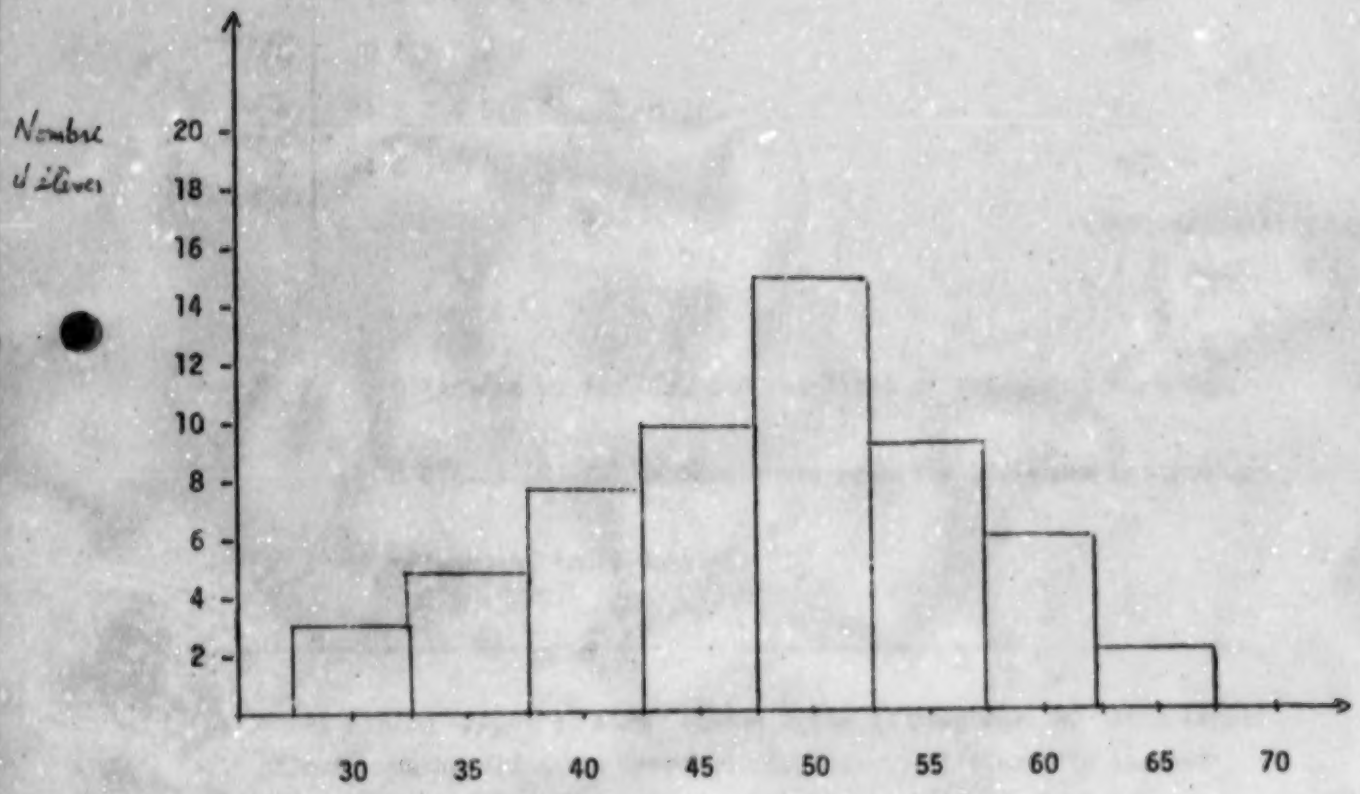
GRAPHE 4 Pictogramme du nombre d'échecs (E) à un examen de statistiques dans six classes de Secondaire V



GRAPHE 5 Pictogramme du nombre d'échecs (☹) à un examen de statistiques dans six classes de Secondaire V

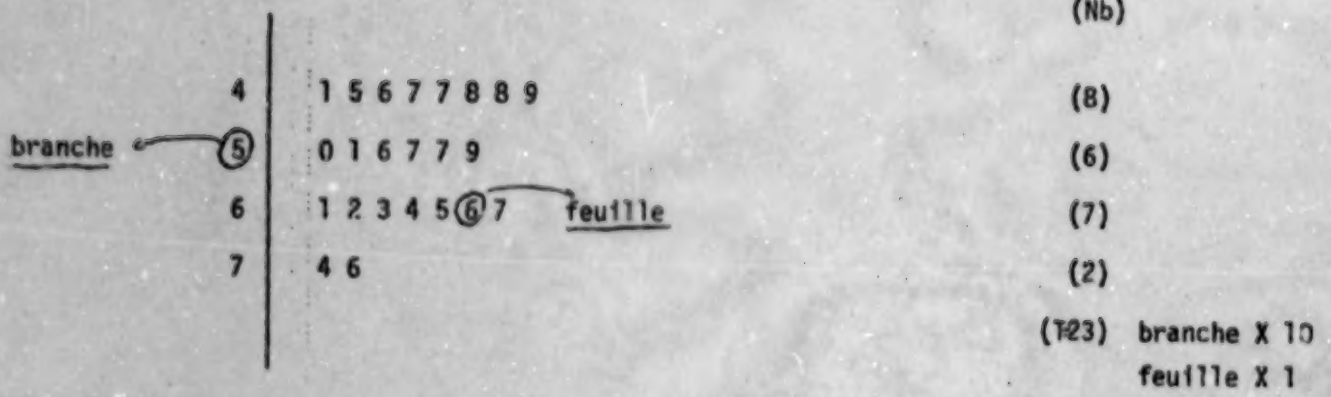


GRAPHE 6 Polygone de fréquences du temps (en secondes) de lecture d'un texte français pour deux classes (58 élèves) de 2ème année du niveau primaire



GRAPHE 7 Histogramme du temps (en secondes) de lecture d'un texte français pour deux classes (58 élèves) de 2ème année du niveau primaire

Chacune des entités numériques à droite de la barre verticale de chaque ligne est nommée feuille ou "leaf".(1)



GRAPHE 8 Diagramme en feuilles des résultats en français d'un groupe d'élèves (N =23) de 6ème année primaire après une année d'enseignement individualisé.(2)

---

(1) Tukey (1977) appelle "stem" chaque ligne du diagramme et "stem label" ce que nous appelons ici branche. Cependant, il identifie souvent (voir pp.9,16...) "stem" et "stem label", ce qui revient à adopter notre définition.

(2) Voir Begin, Y., "Rapport-synthèse de la cinquième année d'évaluation du projet Sage", Document R-143, (INRS-Education, 1981).

Le graphe 8 présente un diagramme en feuilles dans sa plus simple expression. Il est constitué de quatre branches 4, 5, 6 et 7 ayant chacune respectivement 8, 6, 7 et 2 feuilles. Donc en tout 23 feuilles qui correspondent aux 23 données de base 41, 45, 46, 47, 47, 48, 48, 49, 50, 51, 56, 57, 57, 59, 61, 62, 63, 64, 65, 66, 67, 74, 76. En fait, remarquons que dans un diagramme en feuilles, chaque donnée de base est décomposée en deux parties: la branche et la feuille, la même branche pouvant servir pour plusieurs feuilles. Soulignons aussi qu'il est pratique d'ajouter une colonne titrée "Nb" et indiquant le nombre de feuilles par ligne (ou par branche) ainsi que le total (T = 23). De plus, nous avons indiqué la valeur réelle de chaque branche (branche X 10) et de chaque feuille (feuille X 1) dans le coin inférieur droit du diagramme. Mentionnons enfin que, sauf indication contraire, chaque chiffre à droite de la barre verticale sera considéré comme une entité numérique bien distincte rattachée à la branche de la même ligne, donc comme une feuille.

Avant d'introduire un autre type de diagramme dû aussi à Tukey, nous insisterons sur une notion un peu méconnue, la <u>résistance</u> des sommaires numériques, i.e. des points de repère ou indices qui synthétisent l'information contenue dans les données.

Traditionnellement, on insiste surtout sur deux types de sommaire: les <u>mesures de tendance centrale</u> (i.e. les mesures qui donnent une idée du centre de la distribution) et les <u>mesures de dispersion</u> (i.e. les mesures qui donnent une idée de l'étendue de la distribution). Traditionnellement toujours, on présente deux ou trois mesures de tendance centrale et on retient la <u>moyenne arithmétique</u> parce que plus pratique. On utilise rarement les autres parce qu'elles se prêtent moins bien aux calculs. De même, on définit quelques mesures de dispersion et on retient généralement

la variance et sa soeur cadette l'écart-type pour les mêmes raisons "pratiques".

Or la popularité de ces sommaires numériques (moyenne arithmétique, variance ou écart-type) ne doit pas dérouter celui qui pratique l'analyse exploratoire des données. Car il arrive fréquemment, qu'un paquet de données contienne des valeurs aberrantes ce qui risque de donner une idée fausse ou en tout cas fort imprécise de la tendance centrale ou de la dispersion de la distribution. Un simple exemple servira à justifier nos dires.

Considérons les deux paquets de données suivants:

P_1 = } 10,63,64,65,66,68,72,76,79 {

P_2 = } 63,64,65,66,68,72,76,79 {

Une seule donnée "10", sépare P_1 de P_2.

Calculons les sommaires numériques traditionnels. On obtient:

moyenne de $P_1 = \frac{10 + 63 + 64 + 65 + 66 + 68 + 72 + 76 + 79}{9} = 62,6$

variance de $P_1 = \frac{(10-62,6)^2 + (63 - 62,6)^2 + \ldots\ldots + (76 - 62,6)^2 + (79 - 62,6)^2}{8}$

$= 419$

écart-type de $P_1 = \sqrt{\text{variance}} = 20,5$

et moyenne de $P_2 = 69,1$

variance de $P_2 = 35$

écart-type de $P_2 = 5,9$

Deux remarques peuvent dès maintenant être indiquées. Un, la moyenne de P_1 représente très infidèlement la tendance centrale puisque toutes les données sauf une se situent au-dessus de cette valeur. Deux, les importantes fluctuations de la moyenne de 62,6 pour P_1 à 69,1 pour P_2 et de l'écart-type (resp. de la variance) de 20,5 (resp. 419) pour P_1 à 5,9 (resp. 35) pour P_2 sont dues à une seule donnée "10": le moins que l'on puisse dire est que ces sommaires semblent carrément instables et non résistants aux valeurs non conformes ou aberrantes.

Malgré leur popularité, ces mesures semblent donc peu utiles à ce stade-ci (exploratoire) de l'analyse des données (ce qui ne signifie pas qu'elles ne le deviendront pas à un autre stade de l'analyse.) Ce ne sont pas de bons candidats aux points de repère ou sommaires numériques que l'on souhaiterait obtenir en A.E.D. On comprend mieux maintenant l'assertion émise en introduction au présent chapitre et stipulant que les outils utilisés en A.E.D. doivent être résistants aux valeurs aberrantes.

Peut-être devrions-nous investiguer ces sommaires numériques plus résistants parmi les mesures de tendance centrale et de dispersion traditionnellement délaissées. La médiane par exemple ferait peut-être notre affaire. Reprenons les paquets de données P_1 et P_2 et mettons la médiane à l'épreuve. On se souviendra que cette dernière se définit comme le point milieu d'une série de données (en ordre), donc à mi-chemin entre la première et la dernière donnée. C'est-à-dire que la médiane se situe à la position suivante dans la série des données:

$$\frac{\text{position de la 1ère donnée + position de la dernière donnée}}{2}$$

$$= \frac{1 + \text{nombre de données}}{2}$$

Pour nous aider dans les calculs ultérieurs nous avons retranscrit les paquets P_1 et P_2 en indiquant les positions relatives de chaque donnée:

P_1

donnée:	10,	63,	64 ,	65,	66,	68,	72,	76, 79
position:	1,	2,	3,	4,	5,	6,	7,	8, 9

P_2

donnée:	63,	64,	65,	66,	68,	72,	76,	79
position:	1,	2,	3,	4,	5,	6,	7,	8,

On peut maintenant trouver aisément que la position de la médiane de P_1 est $\frac{1 + 9}{2} = 5$. C'est donc "66".

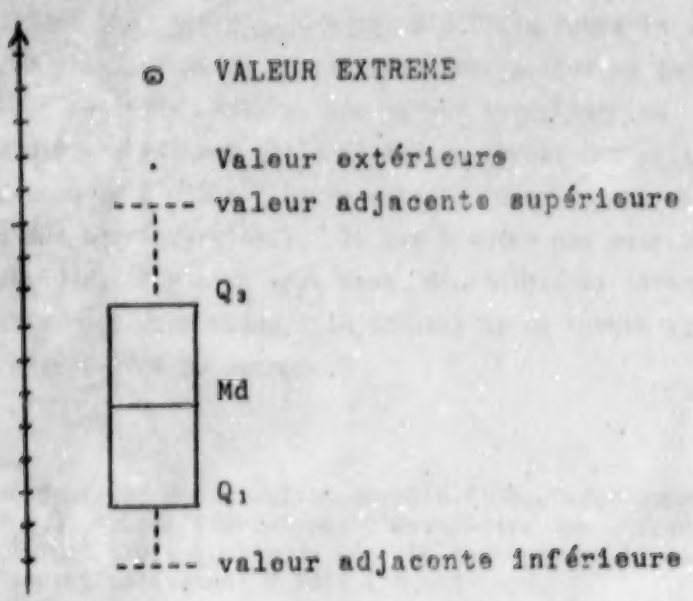
De même la position de la médiane de P_2 est $\frac{1 + 8}{2} = 4,5$

i.e à mi-chemin entre la 4ème et la 5ème donnée. C'est donc $\frac{66 + 68}{2} = 67$,

De la même façon on trouverait que l'étendue interquartile (EI) est une mesure de dispersion résistante aux valeurs aberrantes.

On se souviendra que EI = $Q_3 - Q_1$, où Q_1 est le premier quartile i.e. Q_1 est situé à mi-chemin entre la première valeur du paquet de données (ordonné) et la médiane. De même Q_3 est le troisième quartile, situé à mi-chemin entre la médiane et la dernière valeur du paquet.

Mais ce n'est pas tout d'utiliser des sommaires numériques résistants aux valeurs aberrantes. Encore faut-il pouvoir détecter ces valeurs de façon efficace. Le diagramme en feuilles avec toutes ses vertus n'a pas ce pouvoir. Pour y arriver il faudra introduire un second type de diagramme: le diagramme en boîte.



Comme on peut le voir, ce type de diagramme ressemble à un rectangle, que nous appellerons une boîte, prolongé de chaque côté, de traitillés verticaux et horizontaux que nous appellerons moustaches.

L'axe vertical, à gauche de la boîte, indique que les valeurs représentées dans ce diagramme ont été calculées à l'échelle.

Le premier (Q_1) et troisième quartiles (Q_3) sont situés aux extrémités de la boîte alors que la position de la médiane est représentée par une barre horizontale morcelant la boîte en deux parties (non nécessairement égales!).

En vue de donner la signification des autres valeurs de ce diagramme, nous devrons définir préalablement un certain nombre de termes. Ainsi un saut ou "step" vaut 1,5 fois[1] l'étendue interquartile; une clôture interne ou "inner fence" se situe à la distance d'un saut à l'extérieur de chaque quartile tandis qu'une clôture externe ou "outer fence" se situe à la distance de deux sauts des quartiles; une valeur adjacente se définit comme la valeur du paquet la plus proche de la clôture interne tout en demeurant à l'intérieur de cette clôture; une valeur extérieure se trouve entre la clôture interne et la clôture externe; une valeur extrême se trouve au-delà d'une clôture externe (donc à plus de deux sauts d'un des deux quartiles). Il est à noter que pour chaque paquet de données, il y a un seul saut, deux clôtures internes, deux clôtures externes, deux valeurs adjacentes et un nombre variable de valeurs extérieures ou extrêmes.

---

1 Rappelons que pour une distribution normale EI équivaut approximativement à 1,35 fois l'écart-type; c'est-à-dire que l'écart-type vaut près des trois quarts de EI. En conséquence le saut= 1,5 EI vaut approximativement 2 fois l'écart-type.

Pour terminer cette présentation sommaire, nous donnons un exemple d'un paquet de données que l'on a représenté sous forme de diagramme en feuilles puis sous forme de diagramme en boîte.

Voici d'abord le diagramme en feuilles (graphe 9). Chaque donnée représente le nombre d'étudiants (par 100 000 habitants) engagés dans un programme d'enseignement supérieur pour un des 100 pays considérés.

A l'aide de ce diagramme, on voit qu'il y a un pays obtenant un taux d'étudiants aussi faible que 3 (il s'agit de la Rhodésie) et un autre pays un taux de 1900 étudiants par 100 000 étudiants (bien sûr il s'agit des Etats-Unis).

Il est donc apparent que la variabilité de ces données soit très grande. On remarque aussi une forte concentration des données entre 100 et 390, c'est-à-dire autour de la médiane (Md = 220).

Par ailleurs le diagramme en boîte(graphe 10) offre ici un excellent complément pour la représentation visuelle de ces données puisqu'il nous permet d'un seul coup d'oeil d'accéder à l'étendue totale de la distribution: n'oublions pas, en effet, que ce diagramme est construit à l'échelle.

En outre on peut voir aussi très distinctement l'étendue interquartile (la longueur de la boîte), la position de la médiane, des quartiles et des valeurs adjacentes, mais surtout les valeurs aberrantes (extérieures ou extrêmes). Par exemple on retrouve dans ce cas quatre valeurs aberrantes dont une valeur extrême (les Etats-Unis) et trois valeurs extérieures (Porto-Rico, les Philippines et la Hollande). Une étude analytique pourrait être entreprise concernant ces valeurs. Pourquoi les Etats-Unis sont-ils éloignés des autres pays? Comment se fait-il qu'un pays comme Porto-Rico arrive bon second? S'agit-il d'erreurs ou de faits véritables et si oui comment les expliquer?

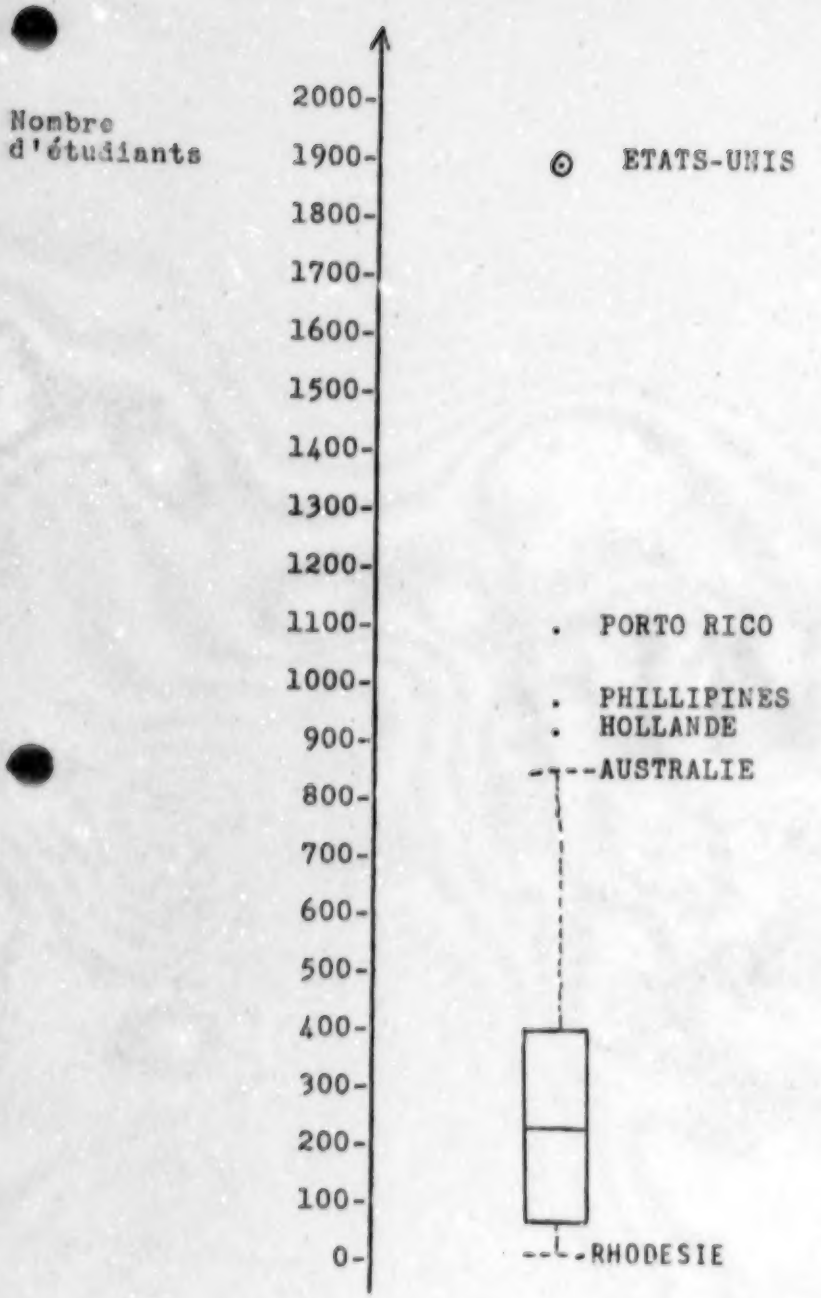
		(Nb)	(Pr)
0*	3 4 4 4 5 5 6 9	(8)	(8)
1	2 4 4 8	(4)	(12)
2	1 4 7 9 9	(5)	(17)
3	4 6	(2)	(19)
4	0 2 9	(3)	(22)
5	6 6	(2)	(24)
Q_1 → 6	1 2 3 4 9	(5)	(29)
7	0 8 8	(3)	(32)
8	3 9	(2)	(34)
9*	0	(1)	(35)
1**	0 1 3 3 4 4 4 6 6 7 7 8 8 9	(14)	(49)
Md → 2	2 2 2 5 5 5 5 5 5 5 5 7 9	(13)	-
Q_3 → 3	2 2 2 4 5 5 6 6 7 9 9 9 9	(13)	(38)
4	0 3 4 5 6 7	(6)	(25)
5	2 2 2 3 3 4 4 7	(8)	(19)
6	4 6 6	(3)	(11)
7	5	(1)	(8)
8	2 3 5	(3)	(7)
9**	2 7	(2)	(4)
1***	1 9	(2)	(2)
2			

(T = 100)

branche X 10 Nb (*)

feuille X 10 Nb (*) - 1

GRAPHE 9 Diagramme en feuilles du nombre d'étudiants (par 100 000 habitants) engagés dans un programme d'enseignement supérieur pour 100 pays



GRAPHE 10 Diagramme en boîte du nombre d'étudiants (par 100 000 habitants) engagés dans un programme d'enseignement supérieur pour 100 pays.